Nucleotide-resolution DNA double-strand breaks mapping by next-generation sequencing

Published in final edited form as:
Nat Methods. 2013 April ; 10(4): 361–365. doi:10.1038/nmeth.2408.

Nucleotide-resolution DNA double-strand breaks mapping by next-generation sequencing

Nicola Crosetto[1,2,11,13,14], **Abhishek Mitra**[3,13], **Maria Joao Silva**[4], **Magda Bienko**[1,2,11], **Norbert Dojer**[3,12], **Qi Wang**[5,6], **Elif Karaca**[5,6], **Roberto Chiarle**[5,6,7], **Magdalena Skrzypczak**[8], **Krzysztof Ginalski**[8], **Philippe Pasero**[4], **Maga Rowicka**[3,9,10,14], and **Ivan Dikic**[1,2,14]

[1]Institute of Biochemistry II, Goethe University Medical School, Frankfurt am Main, Germany [2]Buchmann Institute for Molecular Life Sciences, Goethe University Medical School, Frankfurt am Main, Germany [3]Institute for Translational Sciences, University of Texas Medical Branch at Galveston, Galveston, Texas, USA [4]IGH Institute of Human Genetics, CNRS UPR 1142, Montpellier, France [5]Department of Pathology, Children's Hospital, Boston, Massachusetts, USA [6]Harvard Medical School, Boston, Massachusetts, USA [7]Department of Molecular Biotechnology and Health Sciences, University of Torino, Torino, Italy [8]Laboratory of Bioinformatics and Systems Biology, Centre of New Technologies, University of Warsaw, Warsaw, Poland [9]Department of Biochemistry and Molecular Biology, University of Texas Medical Branch at Galveston, Galveston, Texas, USA [10]Sealy Center for Molecular Medicine, University of Texas Medical Branch at Galveston, Galveston, Texas, USA [12]Institute of Informatics, University of Warsaw, Warsaw, Poland

Abstract

We present a genome-wide method to map DNA double-strand breaks (DSBs) at nucleotide resolution by direct *in situ* breaks labeling, enrichment on streptavidin, and next-generation sequencing (BLESS). We comprehensively validated and tested BLESS using different human and mouse cells, DSBs-inducing agents, and sequencing platforms. BLESS was able to detect telomere ends, Sce endonuclease-induced DSBs, and complex genome-wide DSBs landscapes. As a proof of principle, we characterized the genomic landscape of sensitivity to replication stress in human cells, and identified over two thousand non-uniformly distributed aphidicolin-sensitive regions (ASRs) overrepresented in genes and enriched in satellite repeats. ASRs were also enriched in regions rearranged in human cancers, with many cancer-associated genes exhibiting high sensitivity to replication stress. Our method is suitable for genome-wide mapping of DSBs in various cells and experimental conditions with a specificity and resolution unachievable by current techniques.

[14]Correspondence about computational analyses should be addressed to M.R. (maga.rowicka@utmb.edu), and the rest to N.C. (crosetto@mit.edu) or I.D. (ivan.dikic@biochem2.de) .
[11]Present address: Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[13]Equally contributing authors

ACCESSION CODES The NCBI SRA accession number for all the data presented in this study is SRP018506.

AUTHORS CONTRIBUTIONS N.C. and I.D. conceived and developed BLESS, coordinated the project, and wrote the manuscript. A.M. developed all necessary code and analyzed Illumina data. M.J.S. and P.P. performed ChIP experiments and analysis. M.B. performed microscopy experiments and prepared figures. Q.W., E.K., and R.C. performed Roche 454 experiments and analyzed the data. N.D. contributed to statistical data analysis. M.S. and K.G. performed pair-end Illumina sequencing. M.R. conceived procedures for computational analysis, supervised the analysis, and coordinated the project.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

DNA double-strand breaks (DSBs) can be caused by exogenous and endogenous physical and chemical agents, and appear during apoptosis, meiotic crossing-over, and gene rearrangements [1,2]. Replication fork stalling and collapse also cause DSBs, and are considered the major endogenous source of breaks in cycling cells [3,4]. Unresolved DSBs pose a serious threat to genomic stability, potentially leading to the formation of oncogenic mutations, including translocations, deletions, and amplifications [2]. Despite extensive knowledge on mechanisms of DSBs sensing and repair, the genome-wide landscape of DSBs in different cells and conditions remains largely unknown, mainly because of the lack of methods to map DSBs with high specificity and resolution throughout the genome. Chromatin immunoprecipitation (ChIP) coupled to microarray (ChIP-on-chip) or next-generation sequencing (ChIP-seq) has been applied to map DSBs [5-7]. However, the fact that in ChIP-based methods DSBs are not directly labeled *in situ*, but rather detected indirectly using antibodies targeting specific proteins that bind to DSBs, represents a considerable source of bias. The phosphorylated histone variant H2A.X (γH2A.X) is typically used as a marker of DSBs, but can also mark single-stranded breaks as well the inactive X chromosome [8-10]. Moreover, γH2A.X ChIP signals have been shown to spread tens of kilobases away from a single DSB [11,12], making it difficult to map the exact position of a DSB. Alternatively, the recruitment of replication protein A (RPA) has been used to map DNA damage, but RPA accumulation is partially blocked by 53BP1 [13,14], therefore limiting the use of RPA to cells lacking 53BP1 for reliable results [15]. Other approaches have used capture or direct labeling of single-stranded DNA (ssDNA) followed by microarray analysis, assuming ssDNA is a good proxy for DSBs [16-18]. A bias in these methods, however, is that ssDNA not converted to DSBs (for example, during replication) will yield false positive results. Recently, a method based on breaks labeling with the terminal deoxynucleotydil transferase (TdT) enzyme has been used to detect DSBs at defined locations *in vitro* and in purified genomic DNA from *S. cerevisiae* [19]. This method, however, has not been implemented on a genomic scale, it is not *in situ*, and it does not allow labeling of DSBs with specific barcode sequences, which would be extremely helpful in filtering next-generation sequencing data. Therefore, more resolved and specific genome-wide methods are needed to gain insights in the biology of DNA double-strand breaks in different cell types and conditions. Here, we present a comprehensive experimental and computational approach to directly map DSBs genome-wide, based on direct *in situ* breaks labeling, enrichment on streptavidin, and next-generation sequencing (BLESS).

RESULTS

Method workflow

To avoid specificity issues associated with ChIP or TdT, we devised a protocol for specific and direct *in situ* single-nucleotide resolution labeling and capture of individual DSBs in nuclei purified from mammalian cells (Online Methods). Direct *in situ* labeling prevents labeling of DSBs artificially formed during gDNA extraction, thus minimizing the risk of false positives. After fixation to stabilize chromatin and prevent artificial DSBs, cells are lysed and shortly incubated with proteinase K to purify intact nuclei. As a proof of concept, we optimized this step using both human cell lines as well as mouse tissues from which single-cell suspensions can be easily obtained (Supplementary Fig. 1a). After purification of intact nuclei, DSBs are blunted, 5′-phosphorylated, and finally ligated to a biotinylated linker (proximal) using the highly specific T4 ligase enzyme, which can only ligate double-stranded but not single-stranded breaks. The linker forms a hairpin-like structure, and thus can be ligated to either a blunt DSB or to an identical linker molecule, but cannot form concatemers. The ligatable end of the linker consists of a barcode sequence marking the exact position of ligation followed by the XhoI recognition site. Genomic DNA (gDNA) is extracted and fragmented, and labeled fragments are captured by streptavidin. A second

linker (distal) also containing the XhoI site is attached to the free extremity of captured genomic fragments, enabling PCR amplification and sequencing (Fig. 1a-b). The use of barcoded linkers is a powerful strategy to unequivocally mark DSBs, avoiding background subtraction procedures as in ChIP data analysis.

Method implementation and validation

To implement our DSBs direct labeling and capture procedure, we performed pilot BLESS experiments in HeLa cells and mouse B-lymphocytes, followed by Sanger sequencing and next-generation sequencing on the Roche 454 platform. We performed various controls to exclude substantial false positive labeling due to incomplete washout of proximal linkers, unspecific binding of gDNA to streptavidin beads or mispriming. Only by following the complete BLESS protocol, DNA fragments could be amplified and subjected to sequencing (Supplementary Fig. 1b-d). 88% ± 6.5% (mean ± s.d., $n = 2$) of Roche 454 barcodes reads contained both proximal and distal barcodes, whereas obly 1.5% ± 2% (mean ± s.d., $n = 2$) contained the proximal barcode on both ends (Supplementary Table 1). As a first proof of specificity, we searched for reads mapping in the immunoglobulin heavy chain (IgH) locus among sequences derived from activated mouse B-lymphocytes. Upon B-lymphocyte activation, DSBs are formed in the IgH donor Su region and the downstream acceptor S region, enabling antigen class switch [20]. Accordingly, the density of correctly barcoded reads within these regions was significantly higher than the average read density in the genome (2-fold enrichment, $P = 0.02$, hypergeometric test), even with the relatively modest throughput achievable with the Roche 454 platform.

To increase data throughput, we performed deeper sequencing of BLESS samples using Illumina GAII and HiSeq 2000 platforms (Supplementary Table 1). All sequencing data, including Sanger and Roche 454 sequences, can be accessed at http://www.breakome.eu. In single-end sequencing experiments, the proportions of proximal and distal barcodes among barcoded reads were greatly similar (proximal 52.3% ± 9.8%, and distal 47.7% ± 9.8%, mean ± s.d., $n = 9$). Pair-end sequencing confirmed that over 99% (99.3% ± 0.2%, mean ± s.d., $n = 2$) of BLESS barcoded fragments contained both proximal and distal barcodes, whereas less than 0.8% contained the same barcode on both ends (Fig. 1c and Supplementary Table 1). This result demonstrates that the false positive DSBs labeling rate in BLESS is lower than 1%. We initially deep sequenced HeLa cells – a model system for which a large amount of genome-wide data is available and in which telomeric ends have been well characterized [21]. During BLESS, the 3′ G-overhang of unprotected telomeres – which resembles a DSB repair intermediate [22] – is trimmed down to the first nucleotide of the complementary C-strand, where the biotinylated linker is ligated (Supplementary Fig. 2a). Therefore, we expected accessible telomeric ends to be detected by BLESS. Accordingly, we retrieved telomeric reads derived from the C-strand, with CTAACC being the most frequent (73%) C-strand end, as previously reported [21] (Supplementary Fig. 2b). We also deep sequenced U2OS cells carrying an I-SceI transgenic cassette, after transfecting them with HA-I-SceI (SCE) to induce a single DSB per cell within this cassette (Supplementary Fig. 2c). The density of barcoded reads inside the I-SceI cassette was almost 13,000 fold higher than the average read density in the genome ($P = 6 \times 10^{-300}$, hypergeometric test), further validating our labeling method. These results demonstrate that direct *in situ* labeling of DSBs followed by next-generation sequencing is an effective strategy to identify DSBs at various genomic locations.

Genomic landscape of sensitivity to replication stress

Deep sequencing experiments revealed DSBs sparse throughout the genome, which from now on we refer to as “breakome”. Cells grown in cell culture carry a non negligible amount of DNA breaks caused by a combination of replication stress, physiological apoptosis, and

damage induced by reactive oxygen species [12]. Indeed, even in the absence of any exogenous treatment HeLa cells carried a substantial burden of γH2A.X foci (mean = 3.8, s.d. = 6.7 foci per nucleus, $n = 300$) (Supplementary Fig. 3a-b). In order to single out breaks caused by replication stress, which are believed to be a major source of genome instability [3], we exposed HeLa cells to a dose of aphidicolin – a DNA polymerase inhibitor – that induces replication fork stalling without arresting progression in S-phase [23]. This treatment resulted in significant accumulation of breaks ($P = 10^{-34}$, Kolmogorov-Smirnov test), and increased the amount of labeled DSBs captured by BLESS (Supplementary Fig. 3a-d).

We analyzed aphidicolin sensitivity at various resolutions by comparing deep sequencing data from replicates of samples treated (A) or not (C) with aphidicolin (Supplementary Fig. 4a). We analyzed the data by comparing read numbers within genomic windows with constant mappable length in A vs. C samples, and calculated enrichment P values based on the hypergeometric probability distribution. We computed final Q values based on the Benjamini-Hochberg approach for multiple hypotheses testing [24] (Online Methods). We mapped aphidicolin-sensitive regions (ASRs) at a resolution of 48 kilobases (kb) without and with correction for copy number variation effects due to karyotype and aphidicolin treatment (Supplementary Fig. 4b) which yielded 2,307 and 2,429 significantly correlated ASRs, respectively ($P = 10^{-323}$, hypergeometric test). To calculate the false discovery rate (FDR) related to our approach, we analyzed reads that were mapped to the chromosome Y through sequencing errors (in HeLa cells no reads from chromosome Y are expected). At 48 mappable kb resolution, the calculated FDR was 0.3%. The full list of ASRs and Q values is available at http://www.breakome.eu.

ASRs were non-uniformly distributed along the genome, with an average of 3% of 48 mappable kb regions per chromosome sensitive to aphidicolin, except for chromosomes 5 and 7 which were significantly more sensitive (5%, $P = 10^{-5}$, hypergeometric test) (http://www.breakome.eu). As a comparison, we applied BLESS to HeLa cells treated with neocarzinostatin – a DSBs-inducing drug presumed to yield a more “random” pattern of breaks. Neocarzinostatin-sensitive regions were significantly more uniformly spread along the genome (distance between consecutive sensitive regions: 0.54 ± 1.81 megabases (Mb) and 0.21 ± 0.77 Mb for aphidicolin and neocarzinostatin, respectively, median ± s.d. $P = 10^{-118}$, Kolmogorov-Smirnov test) (Fig. 2a and b). To validate our findings, we compared several top-ASRs with regions displaying no appreciable aphidicolin effect using γH2A.X ChIP. We observed a strong concordance between the aphidicolin effects measured in targeted regions and BLESS results (Fig. 3a-b). Importantly, the genomic locations of ASRs mapped in different experimental replicates were significantly correlated, demonstrating the reproducibility of our method (Supplementary Table 2).

Characterization of aphidicolin-sensitive regions

It has been suggested that repetitive DNA sequences may favor fork stalling and collapse upon replication stress, causing DSBs to appear more frequently at certain genomic regions [3,4]. In particular, repeats prone to form hairpin-like secondary structures might cause the collapse of slowly moving replication forks by directly hindering their progression. To investigate the association between repeats and aphidicolin sensitivity, we applied RepeatMasker [25] to compute the abundance of various DNA repeat families inside ASRs as compared to the rest of the genome (Online Methods). We detected a reproducible strong and significant enrichment in satellites ($P = 5 \times 10^{-137}$) in particular of alpha-type ($P = 8 \times 10^{-198}$, see Online Methods for derivation of P values), a class of repeats that forms hairpin-like secondary structures and is abundant peri/centromeric regions (Fig. 4a, Supplementary Fig. 5, and Supplementary Table 3). Accordingly, ASRs mapped at high resolution (250 mappable nucleotides) were concentrated in peri/centromeric regions (Supplementary Table 4). Another class of repeats, AT dinucleotides, has been associated with a particular group of

genomic regions sensitive to replication stress induced by aphidicolin – common fragile sites (CFSs) [26,27]. While many CFSs were scored as sensitive to aphidicolin following our approach, AT repeats were significantly depleted in ASRs ($P = 10^{-16}$, hypergeometric test).

During replication stress, slowly moving replication forks will have a higher chance of colliding with transcriptional forks [28], resulting in accumulation of DSBs which may be detected by BLESS. Accordingly, upon aphidicolin treatment, we detected a prominent enrichment of DSBs in transcribed regions with the highest enrichment in coding regions ($P = 10^{-10}$, hypergeometric test). We next analyzed genes and ranked them based on the computed probability to develop DSBs anywhere along their length. Top 20% aphidicolin-sensitive genes showed significant enrichment of many gene ontology (GO) terms, particularly related to cell death ($P = 10^{-3}$, hypergeometric test). Gene sensitivity to aphidicolin was significantly associated with gene length ($P = 10^{-18}$, hypergeometric test), in line with the observation that the probability of collisions between replication and transcription forks in active genes seems to increase with gene length [28,29] (Supplementary Fig. 6).

Finally, we investigated if ASRs mapped by BLESS are also associated with genomic regions or genes frequently rearranged in human cancers. Replication stress-driven genomic instability has been observed in many tumors, where it is thought to be an important cause of cancer rearrangements [30,31]. We used data from a cohort of over 2,700 human cancers [32], and found a modest, but significant enrichment of regions displaying amplifications or deletions inside ASRs as compared to the rest of the genome ($P = 0.005$, derived as described in Online Methods) (Fig. 4b). We next compared aphidicolin-sensitive genes with the Cancer Gene Census [33,34], a collection of over 400 well annotated cancer genes, the majority of which is involved in translocations. Cancer genes were more likely to overlap with ASRs than non-cancer genes ($P = 0.04$, hypergeometric test), and the fraction of genes with 5′ end in 2 Mb vicinity of the center of a 48 mappable kb ASR, and containing that ASR center inside, was higher for cancer genes than others ($P = 0.02$, hypergeometric test) (Fig. 4c). Among most aphidicolin-sensitive genes, cancer genes were overrepresented ($P = 0.04$, hypergeometric test), including prominent oncogenes like *EGFR*, *MET*, *ABL1*, and *MLL*, which are typically mutated by translocation or amplification (Fig. 4d). The full list of genes ranked by aphidicolin sensitivity, and GO analysis results are available at http://www.breakome.eu.

DISCUSSION

DNA double-strand breaks represent a major threat to genomic stability, and understanding the sensitivity of the genome to various DNA insults will be instrumental to implement effective preventive and treatment strategies. Chronic exposure to ionizing radiation is common among several professionals, including medical radiation personnel, pilots and flight attendants, as well as cosmonauts. Moreover, replication errors and reactive oxygen species generated as by- products of metabolism have been estimated to cause breaks at a frequency as high as fifty DSBs per cell per day [12]. In spite of this pervasive threat, our knowledge on how the genome breaks in response to various insults as well as our technology to reliably detect DSBs are still in their infancy. In BLESS, DSBs are directly *in situ* labeled with the highly specific enzyme T4 ligase with biotinylated oligonucleotides that carry a defined barcode sequence. Unlike empirical background subtraction procedures employed in ChIP-based methods to account for non-specific binding, direct DSBs labeling in BLESS ensures high specificity of breaks detection, which can be then unambiguously identified by the presence of barcode sequences. Another important advantage of BLESS over ChIP-based methods is the ability to directly mark DSBs at nucleotide resolution *in situ*, rather then relying on proxies such as γH2A.X which can be found thousand kilobases

away from the actual original DSB [11]. It should be noted that – at least for DSBs repaired by homologous recombination – labeling can occur away from the initial breakpoint due to 5′ end resection. This property can be exploited by inducing DSBs at known genomic positions (analogous to U2OS cells used in this study), obtaining a zoom-in view into the kinetics of DSBs repair *in vivo*, and how this is influenced by the genomic context.

Our method is general and organism-independent, providing genome-wide maps of DSBs for multiple cell types and conditions. Our computational methods and software tools allow to obtain and analyze high-confidence genome-wide DSBs maps, and to account for copy number variation effects attributable to the karyotype of cells analyzed and/or the effects of the treatment used to induce DSBs. Our results demonstrate that hypothesis-driven feature analysis of genomic regions identified by BLESS as sensitive to a specific treatment can help explore the basis of genomic instability at a genome-wide level. In the future, our method could be combined with ultra-deep sequencing of selected regions enriched, for example, by exome capture [35] or reduced representation sequencing [36], thus providing a high-definition picture of the sensitivity of specific regions to DSBs-inducing agents. Finally, the design principle of BLESS could also be exploited for *in situ* DSBs labeling and visualization by super-resolution microscopy. The labeling method and the computational approaches described here represent a valuable resource for the DNA damage research community, providing tools to map and analyze breakomes in a variety of organisms and conditions with a precision and resolution currently unattainable with other methodologies.

ONLINE METHODS

Cells, reagents, and immunocytofluorescence

To obtain primary mouse single-cell suspensions, we squeezed testes and spleens from C57BL6/J mice between two microscope slides in a Petri dish filled with trypsin. We flushed bone marrow out of femurs and tibias from the same animals using a syringe filled with trypsin. We purified and activated B-lymphocytes as previously described [37]. Prior to fixation for BLESS, we removed dead cells using a Ficoll gradient. We filtered cell suspensions through MACS Pre-Separation filters (Milteniybiotec), and then fixed them according to the BLESS protocol. We obtained IMR90 primary fibroblasts and HeLa cells from ATCC, and U2OS_DRH-1 cells from Y. Shiloh (Tel Aviv University). Andrew J. Pierce (University of Kentucky) kindly communicated details on the construction of U2OS_DRH-1 cells. We transfected pcBAS-I-SceI and pCAGGs plasmids (kindly donated by Y. Shiloh) into U2OS_DRH-1 cells using Fugene (Roche) following the manufacturer's instructions. After BLESS, we cloned gDNA fragments into pEGFP-C1 (BD Biosciences). We applied aphidicolin (Sigma) onto cells at 0.4 μM for 18 h, and neocarzinostatin (Sigma) at 200 ng/ml for 45 min. We obtained oligonucleotide linkers from Sigma and annealed them in 1× T4 ligase buffer (NEB). Linkers and primers used are listed in Supplementary Table 5. We visualized γH2A.X foci by immunocytofluorescence (Millipore # 05-636), and counted them as previously described [38].

Breaks Labeling, Enrichment on Streptavidin, and Sequencing (BLESS)

A detailed, step-by-step protocol to perform BLESS can be found in the supporting webpage http://www.breakome.eu. Briefly, to prepare purified nuclei for *in situ* ligation, we fixed five million cells as single-cell suspensions in growth medium with 2% formaldehyde for 30 min at room temperature, and then washed them one time in ice-cold 1× PBS. To prepare single-nucleus suspensions, we first lysed fixed cells in a buffer containing 10 mM Tris-HCl, 10 mM NaCl, 1 mM EDTA, 1 mM EGTA, 0.2% NP-40 pH 8 for 90 min at 4 °C, and then in a buffer containing 10 mM Tris-HCl, 150 mM NaCl, 1 mM EDTA, 1 mM EGTA, 0.3% SDS pH 8 for 45 min at 37 °C. We resuspended lysed cells in 1× NEBuffer 2 (NEB)

supplemented with 0.1% Triton X-100 and Proteinase K at 100 μg/ml final concentration. We mildly rotated cells for a short time at 37 °C (8 min for HeLa, 4 min for mouse B-lymphocytes), after which we transferred them onto ice. We quenched proteinase K by adding an equal volume of buffer supplemented with PMSF.

We washed purified nuclei twice in 1× NEBuffer 2 supplemented with 0.1% Triton X-100, and then once in blunting buffer (NEB) supplemented with 100 μg/ml BSA. We performed blunting using the Quick Blunting kit (NEB) according to the manufacturer's instructions in a final volume of 100 μl, for 45 min at room temperature. Afterwards, we washed nuclei twice in 1× NEBuffer 2 supplemented with 0.1% Triton X-100, once in 1× T4 ligase buffer supplemented with 0.1% Triton X-100, and once in 1× T4 ligase buffer. We performed *in situ* ligation for 18-20 h at 16 °C in 25 μl final volume using 1.5 μl of T4 ligase (NEB) and 5 μl of 10 μM proximal linker previously annealed in 1× ligase buffer. After ligation, we washed nuclei three times ina high-salt buffer (W&B) containing 5 mM Tris-HCl, 1 mM EDTA, 1 M NaCl pH 7.5 supplemented with 0.1% Triton X-100. Afterwards, we extracted gDNA by incubating nuclei in 1× NEBuffer 2 with 0.5% Triton X-100 and Proteinase K at 200 μg/ml final concentration for one hour at 65 °C shaking, followed by isopropanol-ethanol purification. We fragmented purified gDNA for 18-20 h at 16 °C using 6 U of HaeIII (NEB) every million cells fixed.

To capture labeled DSBs, we rotated 20 μg of purified gDNA with 5μl of Dynabeads MyOne C1 (Invitrogen) in W&B buffer supplemented with 0.1% Triton X-100 for 30 min at +4 °C. Afterwards, we washed beads three times in W&B buffer supplemented with 0.1% Triton X-100, and them resuspended in 37 μl of 1× T4 ligase buffer. We added 10 μl of distal linker previously annealed at 10 μM in 1× ligase buffer, and 3 μl of T4 ligase to the beads, and the reaction was carried on for 16-18 h at 16 °C. After distal linker ligation, we washed beads twice in W&B buffer supplemented with 0.1% Triton X-100 at room temperature, and then digested captured fragments with I-Sce (NEB) in 25 μl final volume for 4 h at 37 °C. Afterwards, we centrifuged beads, and stored the supernatant at 20 °C until PCR.

We used all the volume of supernatant after I-Sce digestion to prepare multiple PCR amplification reactions (5 μl per reaction) using Phusion polymerase (NEB) and the appropriate primers pair depending on the downstream sequencing platform (Supplementary Table 5). We performed eighteen amplification cycles using conditions recommended by the manufacturer and T_a = 55 °C. To remove unused primers, we purified PCR products in gel using the DNA Gel Extraction kit (Qiagen). Before Illumina library preparation, we digested purified PCR products with XhoI (NEB) to cleave terminal I-SceI sequences derived from linkers, and again gel-purified them.

Next-generation sequencing

Sequencing was either outsourced (imaGenes GmbH, Berlin, Germany, and ServiceXS, Leiden, Netherlands) or performed in-house as summarized in Supplementary Table 1. We prepared samples for Roche 454 sequencing using indexing barcodes-containing primers during the PCR step in BLESS (Supplementary Table 5). We purified PCR products of size comprised between 300 and 800 nt in gel, and analyzed them on the 2100 Bioanalyzer (Agilent) prior to sequencing. For BLESS Illumina library preparation, we used the TruSeq DNA sample preparation kit v2 (Illumina) without DNA fragmentation and library size selection. For gDNA sequencing, we sheared gDNA with Covaris S220 AFA (Covaris) according to the manufacturer's instructions prior to Illumina library preparation. We assessed library quality and quantity on the 2100 Bioanalyzer (Agilent) using the High Sensitivity DNA Kit (Agilent), and by qPCR with the Kapa Library quantification Kit (KapaBiosystems). We generated clusters on the Illumina flow cell using the automatic cBot

station and the TruSeq PE Cluster Kit v3-cBot-HS. We carried out sequencing by synthesis on Illumina HiSeq 2000 system using the TruSeq SBS Kit v3-HS chemistry.

ChIP and qPCR

We performed ChIP assays as previously described [39], with minor modifications. We purified immuno-precipitated and input DNA with phenol-chloroform, and analyzed it by real-time qPCR using primers listed in Supplementary Table 6. We compared the amount of DNA captured in untreated vs. aphidicolin-treated HeLa cells by qPCR followed by data analysis according to the ΔC_t method [40], using the C_t values obtained for each primer pair in sample C1 as a reference for the C_t values obtained for the same primer pair in sample A1. We used three technical replicates for each sample. Primers are listed in Supplementary Table 7.

Computational analyses

Roche 454 data—We analyzed Roche 454 data using previously described scripts developed in the Chiarle lab [37] adapted for BLESS linkers. Briefly, we aligned sequences to the mouse reference genome (GRCm38/mm10) using BLAT, and subsequently filtered them for the BLESS proximal linker using BLAST. We further processed filtered reads to remove PCR repeats (including repeats slightly divergent due to sequencing errors), invalid alignments (including alignment scores < 30, reads with multiple alignments having a score difference <4, and alignments having ≤ 10 nt gaps), and linker ligation artifacts (for example, random HaeIII restriction sites ligated to proximal linker).

Illumina data—We analyzed Illumina data using the Instant-seq software suite developed in the Rowicka Lab. We filtered reads from fastq files requiring a Phred score ≥ 20 for every base, and trimmed them at the point where the Phred score of an examined base fell below 20. We retained all reads with length ≥ 34 nt as high-quality filtered reads, and scanned them for the presence of the exact proximal or distal barcode. After removal of barcodes, we aligned reads ≥ 23 nt to the GRCh37/hg19 assembly of the human genome. We only retained sequences mapping without mismatches to unique (U0) or multiple (R0) positions.

To identify aphidicolin-sensitive regions (ASRs), we compared the number of reads in A (aphidicolin-treated) and C (untreated) samples using windows with constant number of mappable bases to account for the variation in mappability along the human genome, and produce more statistically robust comparisons of the number of reads between different windows. As a comparison, we also used windows of constant length, demonstrating that our approach is not biased towards detecting ASRs in repetitive regions (Supplementary Fig. 7a-b). Using a pre-computed mappability map corresponding to 45 nt reads, we moved windows of chosen mappable length across each chromosome, and calculated enrichment *P* values based on the hypergeometric probability distribution. The parameters for calculating hypergeometric *P* values were the following: (i) total number of mapped reads in experiments A and C; (ii) total number of mapped reads in experiment A; (iii) total number of mapped reads in the sliding window in experiments A and C; (iv) total number of mapped reads in the sliding window in experiment A only. We computed final *Q* values using the Benjamini-Hochberg correction for multiple hypotheses testing [24]. We first analyzed individual sample replicates (A1 vs. C1; A2 vs. C2; A3 vs. C3; A4 vs. C4), and found that ASRs were highly correlated (Supplementary Table 2). Therefore, for all subsequent analyses, we pooled samples so that C = (C1 + C2 + C3 + C4) and A = (A1 + A2 + A3 + A4). To account for copy number variation effects due to karyotype or aphidicolin treatment, we sequenced gDNA derived from normal human fibroblasts (g-F), HeLa (g-C), and HeLa treated with aphidicolin (g-A), and mapped it to the human reference genome (GRCh37/hg19) (Supplementary Table 1). Using windows with constant number of

mappable bases, we first calculated the expected number of reads in each window based on the g-F sample, and then compared it to normalized reads obtained from g-C and g-A samples. If the ratio of reads was higher or lower than the expected number, we computed the *P* value of the corresponding enrichment or depletion. *P* values were then corrected for multiple hypotheses testing using Benjamini-Hochberg correction [24], and significantly enriched windows ($P \leq 0.05$) were finally annotated as high-confidence, karyotype- and aphidicolin-corrected ASRs. We applied the same procedure to identify neocarzinostatin-sensitive regions (NSRs).

To compute gene sensitivity to aphidicolin, we obtained the start and end coordinates of each gene annotated in GencodeV12 by linearly combining all annotated transcripts. We analyzed each gene by using the same approach described above for ASRs, using the start-end coordinates of the gene as the start-end coordinates of genomic windows. Genes that did not have any mappable bases within the boundaries of the window were not analyzed. We inferred the relationship between aphidicolin sensitivity and gene length by computing rank correlation. We also computed the overlap between genes and ASRs, by agglomerating results from fifty different ASR maps with resolutions in the range of 10-60 mappable kb. For each window length, we first calculated the percentage overlap of a gene with ASRs, and then averaged percentages over all lengths. To assess the significance of overlap enrichment, we computed the distribution of averaged overlap for each gene. We calculated the exact distribution under the assumption that the aphidicolin sensitivity of each window is independently assigned, with probability depending on its length (a different percentage of windows was aphidicolin-sensitive for different window sizes as shown in Supplementary Fig. 4a). We computed the exact distribution step-by-step for consecutive window lengths using dynamic programming. Based on this distribution, we calculated for each gene the *P* value of its overlap enrichment, and then applied Benjamini-Hochberg multiple hypotheses testing correction [24] to the whole gene lists. We compared ASRs enrichment in cancer-associated genes vs. all genes using both the Kolmogorov-Smirnov and hypergeometric test, obtaining the same result ($P = 0.04$). Finally, we identified genes that have transcripts with 5′ end within 2 Mb vicinity of an ASR center using the Gencode annotation of genes. We extracted locations of all transcripts, and for each ASR we identified transcripts in its proximity, and calculated the distance from the center of the ASR to the transcription start site. We analyzed the list of transcripts within 2 Mb distance of an ASR, and reported the corresponding gene along with the exact distance and ASR center location (within or outside). This list was binned at various intervals of distances. Finally, we analyzed the list of transcripts in vicinity of ASRs and outputted a list of gene groups binned according to the distance from the center of the ASR to the closest transcription start site (Fig. 4c). We computed statistics for differences between groups using the hypergeometric probability distribution.

For biological characterization of ASRs, we created feature datasets from RepeatMasker [25] and from the CpG Islands tracks of UCSC Genome Browser [41] and genome-wide summary data of Tumorscape portal [42]. We obtained cancer-associated genes from the Sanger Institute's Cancer Gene Census [33,34]. When required, we mapped genome coordinates to the GRCh37/hg19 assembly of the human genome using liftover. As repetitive regions tend to be underrepresented among uniquely mapped reads, we corrected for differences in mappability. To determine if a given genomic feature is enriched in ASRs, we computed the proportion of mappable nucleotides belonging to both ASRs and the feature, as well as the proportion of ASRs among all the intervals considered. Next, we performed 100,000 permutations of ASR assignments among the windows considered. Based on these permutations, we calculated the empirical distribution of the ratio under the null hypothesis that the given feature and ASRs are independently distributed in the human genome. We used this distribution to estimate the *P* value for the feature enrichment inside ASRs. This

method yielded a *P* value resolution of 10^{-5}, which was too low for features particularly well correlated with fragility (Supplementary Table 3). In such cases, we analyzed empirical ratio distributions. We observed that the normal distribution fits well for long windows (48 mappable kb). For short windows (250 and 2,000 mappable nt), the number of ASR mappable nucleotides within featured regions was almost always the integer multiple of the window length, and multiplication factors followed the geometric distribution. Thus, we analytically determined *P* values according to these distributions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Y. Shiloh (Tel Aviv University) and Andrew J. Pierce (University of Kentucky) are acknowledged for kindly providing U2OS_DRH-1 cells and I-Sce plasmids. We are grateful to T. Włodarski, A. R. Lehmann, G. Fudenberg, and A. Kudlicki for insightful discussions, critical reading of the manuscript, and help with data analysis. This work was supported by grants from Deutsche Forschungsgemeinschaft, the Cluster of Excellence “Macromolecular Complexes” of the Goethe University Frankfurt (EXC115), the LOEWE funded OSF network, LOEWE Gene and Cell Therapy Center and the European Research Council under the European Union’s Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° [250241-LineUb] to I.D.; by grants from the Associazione Italiana per la Ricerca sul cancro (AIRC), from the International Association for Cancer Research (AICR) and grant FP7 ERC-2009- StG (Proposal No. 242965 -- “Lunely”) to R.C.; by grants from the Foundation for Polish Science (TEAM), the National Science Centre (2011/02/A/NZ2/00014), and the European Regional Development Fund under Innovative Economy Programme (POIG.02.02.00-14-024/08-00) to K.G.; by grants from Ligue contre le Cancer (équipe labellisée), ANR (RepliCare), and INCa to P.P.; and by grant 1UL1RR029876-01 from the National Center for Research Resources, National Institutes of Health, and ITS UTMB “Novel Methods” grants to M.R. M.B. is a recipient of a Human Frontiers Science Program Long Term Fellowship.

REFERENCES

1. Paigen K, Petkov P. Mammalian recombination hot spots: properties, control and evolution. Nat Rev Genet. 2010; 11:221–233. [PubMed: 20168297]
2. Jackson SP, Bartek J. The DNA-damage response in human biology and disease. Nature. 2009; 461:1071–1078. [PubMed: 19847258]
3. Branzei D, Foiani M. Maintaining genome stability at the replication fork. Nature Reviews Molecular Cell Biology. 2010; 11:208–219.
4. Branzei D, Foiani M. The DNA damage response during DNA replication. Current Opinion in Cell Biology. 2005; 17:568–575. [PubMed: 16226452]
5. Szilard RK, et al. Systematic identification of fragile sites via genome-wide location analysis of gamma-H2AX. Nat. Struct. Mol. Biol. 2010; 17:299–305. [PubMed: 20139982]
6. Harrigan JA, et al. Replication stress induces 53BP1-containing OPT domains in G1 cells. J. Cell Biol. 2011; 193:97–108. [PubMed: 21444690]
7. Seo J, et al. Genome-wide profiles of H2AX and -H2AX differentiate endogenous and exogenous DNA damage hotspots in human cells. Nucleic Acids Research. 2012 doi:10.1093/nar/gks287.
8. Marti TM, Hefner E, Feeney L, Natale V, Cleaver JE. H2AX phosphorylation within the G1 phase after UV irradiation depends on nucleotide excision repair and not DNA double-strand breaks. Proc. Natl. Acad. Sci. U.S.A. 2006; 103:9891–9896. [PubMed: 16788066]
9. Tuduri S, et al. Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription. Nat. Cell Biol. 2009; 11:1315–1324. [PubMed: 19838172]
10. Chadwick BP, Lane TF. BRCA1 associates with the inactive X chromosome in late S-phase, coupled with transient H2AX phosphorylation. Chromosoma. 2005; 114:432–439. [PubMed: 16240122]
11. Iacovoni JS, et al. High-resolution profiling of gammaH2AX around DNA double strand breaks in the mammalian genome. EMBO J. 2010; 29:1446–1457. [PubMed: 20360682]
12. Bonner WM, et al. γH2AX and cancer. Nature Publishing Group. 2008; 8:957–967.

13. Bunting SF, et al. 53BP1 inhibits homologous recombination in Brca1-deficient cells by blocking resection of DNA breaks. Cell. 2010; 141:243–254. [PubMed: 20362325]
14. Bothmer A, et al. 53BP1 regulates DNA resection and the choice between classical and alternative end joining during class switch recombination. J. Exp. Med. 2010; 207:855–865. [PubMed: 20368578]
15. Hakim O, et al. DNA damage defines sites of recurrent chromosomal translocations in B lymphocytes. Nature. 2012; 484:69–74. [PubMed: 22314321]
16. Blitzblau HG, Bell GW, Rodriguez J, Bell SP, Hochwagen A. Mapping of meiotic single-stranded DNA reveals double-stranded-break hotspots near centromeres and telomeres. Curr. Biol. 2007; 17:2003–2012. [PubMed: 18060788]
17. Feng W, et al. Genomic mapping of single-stranded DNA in hydroxyurea-challenged yeasts identifies origins of replication. Nat. Cell Biol. 2006; 8:148–155. [PubMed: 16429127]
18. Feng W, Bachant J, Collingwood D, Raghuraman MK, Brewer BJ. Centromere replication timing determines different forms of genomic instability in Saccharomyces cerevisiae checkpoint mutants during replication stress. Genetics. 2009; 183:1249–1260. [PubMed: 19805819]
19. Leduc F, et al. Genome-wide mapping of DNA strand breaks. PLoS ONE. 2011; 6:e17353. [PubMed: 21364894]
20. Dudley DD, Chaudhuri J, Bassing CH, Alt FW. Mechanism and control of V(D)J recombination versus class switch recombination: similarities and differences. Adv. Immunol. 2005; 86:43–112. [PubMed: 15705419]
21. Sfeir AJ, Chai W, Shay JW, Wright WE. Telomere-end processing the terminal nucleotides of human chromosomes. MOLCEL. 2005; 18:131–138.
22. Palm W, de Lange T. How shelterin protects mammalian telomeres. Annu. Rev. Genet. 2008; 42:301–334. [PubMed: 18680434]
23. Casper AM, Nghiem P, Arlt MF, Glover TW. ATR regulates fragile site stability. Cell. 2002; 111:779–789. [PubMed: 12526805]
24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B. 1995
25. Smit, A.; Hubley, R. RepeatMasker Open-3.0. 1996–2004. Institute for Systems Biology; 2004.
26. Durkin SG, Glover TW. Chromosome fragile sites. Annu. Rev. Genet. 2007; 41:169–192. [PubMed: 17608616]
27. Zhang H, Freudenreich CH. An AT-rich sequence in human common fragile site FRA16D causes fork stalling and chromosome breakage in S. cerevisiae. MOLCEL. 2007; 27:367–379.
28. Kim N, Jinks-Robertson S. Transcription as a source of genome instability. Nat Rev Genet. 2012; 13:204–214. [PubMed: 22330764]
29. Helmrich A, Ballarino M, Tora L. Collisions between Replication and Transcription Complexes Cause Common Fragile Site Instability at the Longest Human Genes. MOLCEL. 2011; 44:966–977.
30. Halazonetis TD, Gorgoulis VG, Bartek J. An oncogene-induced DNA damage model for cancer development. Science. 2008; 319:1352–1355. [PubMed: 18323444]
31. Negrini S, Gorgoulis VG, Halazonetis TD. Genomic instability--an evolving hallmark of cancer. Nature Reviews Molecular Cell Biology. 2010; 11:220–228.
32. De S, Michor F. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. Nat. Biotechnol. 2011; 29:1103–1108. [PubMed: 22101487]
33. Futreal PA, et al. A census of human cancer genes. Nat. Rev. Cancer. 2004; 4:177–183. [PubMed: 14993899]
34. Santarius T, Shipley J, Brewer D, Stratton MR, Cooper CS. A census of amplified and overexpressed human cancer genes. Nature Publishing Group. 2010; 10:59–64.
35. Ng SB, Turner EH, Robertson PD, Flygare SD. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009
36. Altshuler D, et al. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature. 2000; 407:513–516. [PubMed: 11029002]

37. Chiarle R, et al. Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. Cell. 2011; 147:107–119. [PubMed: 21962511]

38. Crosetto N, et al. Human Wrnip1 is localized in replication factories in a ubiquitin-binding zinc finger-dependent manner. J. Biol. Chem. 2008; 283:35173–35185. [PubMed: 18842586]

39. Tyteca S, Vandromme M, Legube G, Chevillard-Briet M, Trouche D. Tip60 and p400 are both required for UV-induced apoptosis but play antagonistic roles in cell cycle progression. EMBO J. 2006; 25:1680–1689. [PubMed: 16601686]

40. Schmittgen TD, Livak KJ. Analyzing real-time PCR data by the comparative CT method. Nature protocols. 2008

41. Fujita PA, et al. The UCSC Genome Browser database: update 2011. Nucleic Acids Research. 2011; 39:D876–82. [PubMed: 20959295]

42. Beroukhim R, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010; 463:899–905. [PubMed: 20164920]

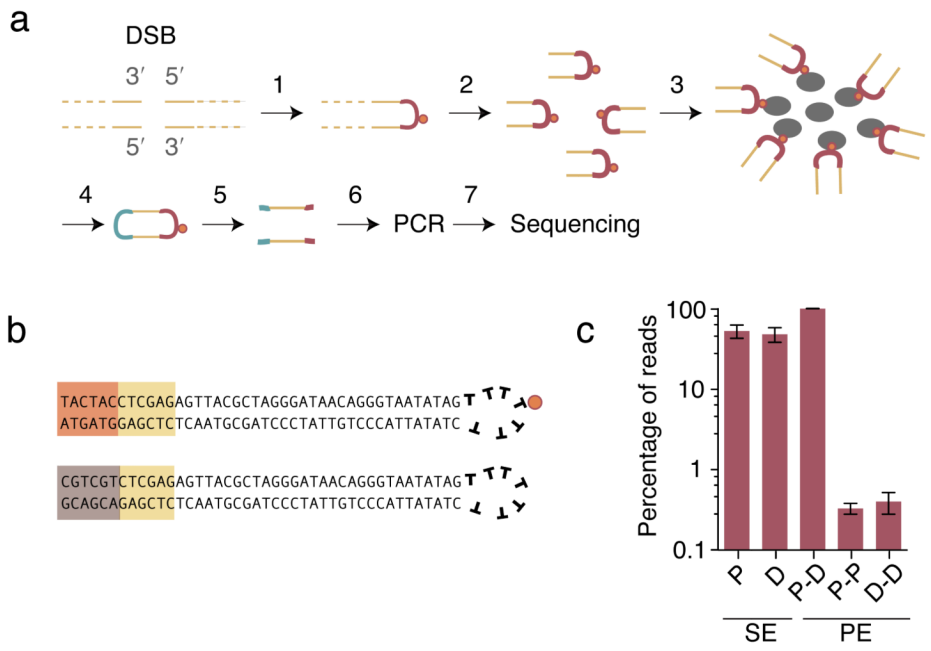


Figure 1.

BLESS workflow and specificity. (**a**) DSBs are ligated *in situ* to a proximal linker (red arch) covalently linked to biotin (orange oval) (1), gDNA is extracted and fragmented (2), and labeled fragments are captured on streptavidin beads (gray ovals) (3). A distal linker (blue arch) is then ligated to the free extremity of captured fragments (4), and fragments are released by linker digestion with I-SceI (5). Released fragments are amplified by PCR using linker-specific primers (6), and sequenced (7). (**b**) Structure of linkers. Both proximal (P) and distal (D) linkers share an XhoI site (yellow), the I-SceI endonuclease minimal recognition site (non-highlighted letters), and a seven-thymine loop (bold). Each linker contains a specific barcode sequence marking the ligation site (orange and brown). The proximal linker is biotinylated (orange oval). (**c**) Proportion of fragments with proximal (P) and distal (D) barcodes in single-end (SE) and pair-end (PE) Illumina sequencing experiments. Mean ± s.d. is shown.

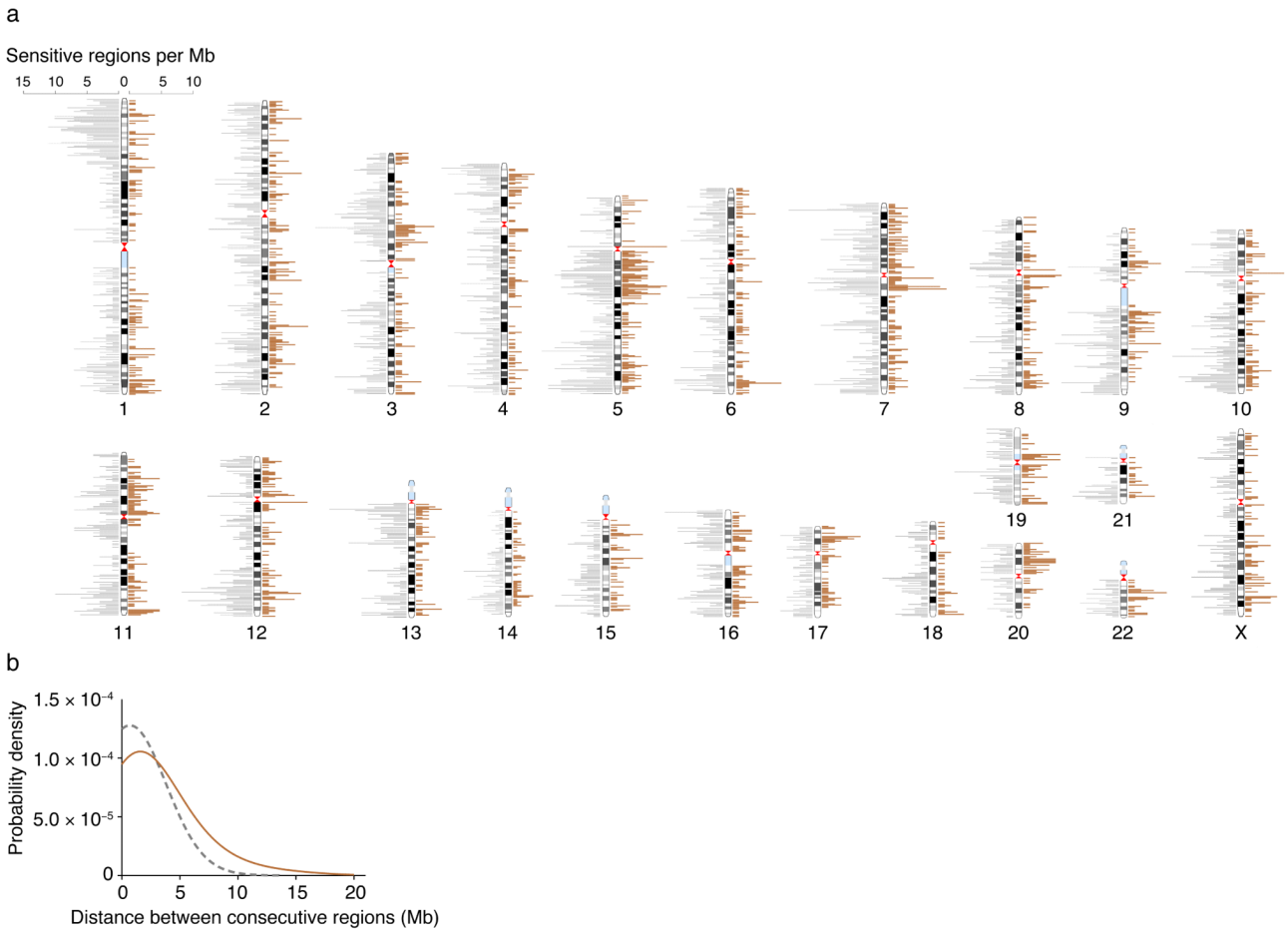
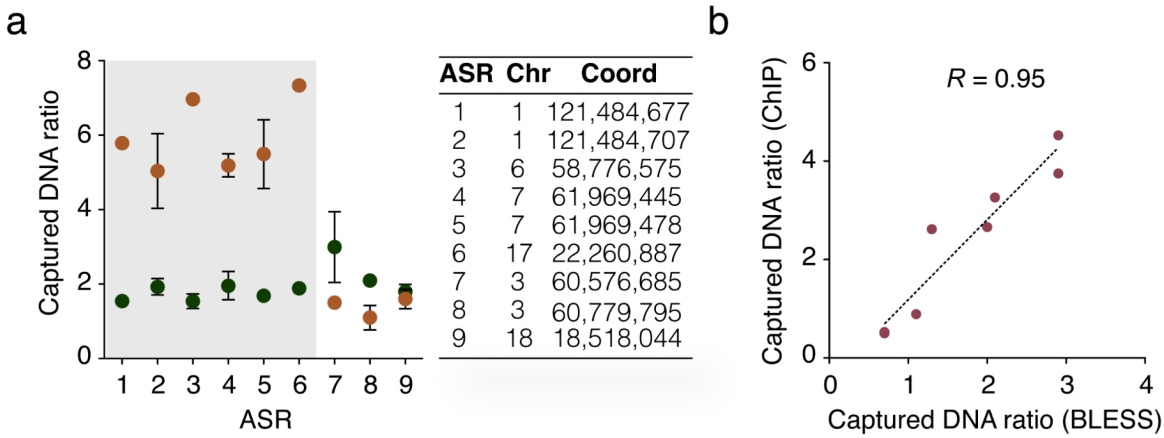


Figure 2.

Example of HeLa breakomes associated with specific treatments. (**a**) Genome-wide aphidicolin (orange) and neocarzinostatin (gray) sensitivity landscapes in HeLa cells, corrected for karyotype and aphidicolin effects. Bars represent the density per one Mb bin of 48 mappable kb ASRs corrected for copy number variation effects. Individual regions and mappability maps are shown in detail in Supplementary Fig. 4b, since non-mappability can artificially lower the number of significant regions per Mb. (**b**) Frequency distribution of genomic distances between the centers of consecutive aphidicolin- and neocarzinostatin-sensitive regions.



ASR	Chr	Coord
1	1	121,484,677
2	1	121,484,707
3	6	58,776,575
4	7	61,969,445
5	7	61,969,478
6	17	22,260,887
7	3	60,576,685
8	3	60,779,795
9	18	18,518,044

Figure 3.

ASRs validation. (**a**) Fraction of input DNA captured by ChIP in regions with significant (grey highlight) vs. non-significant aphidicolin effect in HeLa cells treated (orange) or not (green) with aphidicolin. Mean ± s.d. are shown for $n = 3$ biological replicates. Genomic coordinates of amplicons analyzed by qPCR are reported. Chr: chromosome. Coord: genomic coordinate. (**b**) Comparison of aphidicolin effect measured by BLESS vs. ChIP in regions described in (**a**). Captured DNA ratio: ratio of captured DNA in aphidicolin-treated (A) vs. control (C) HeLa. R: Pearson's correlation coefficient.

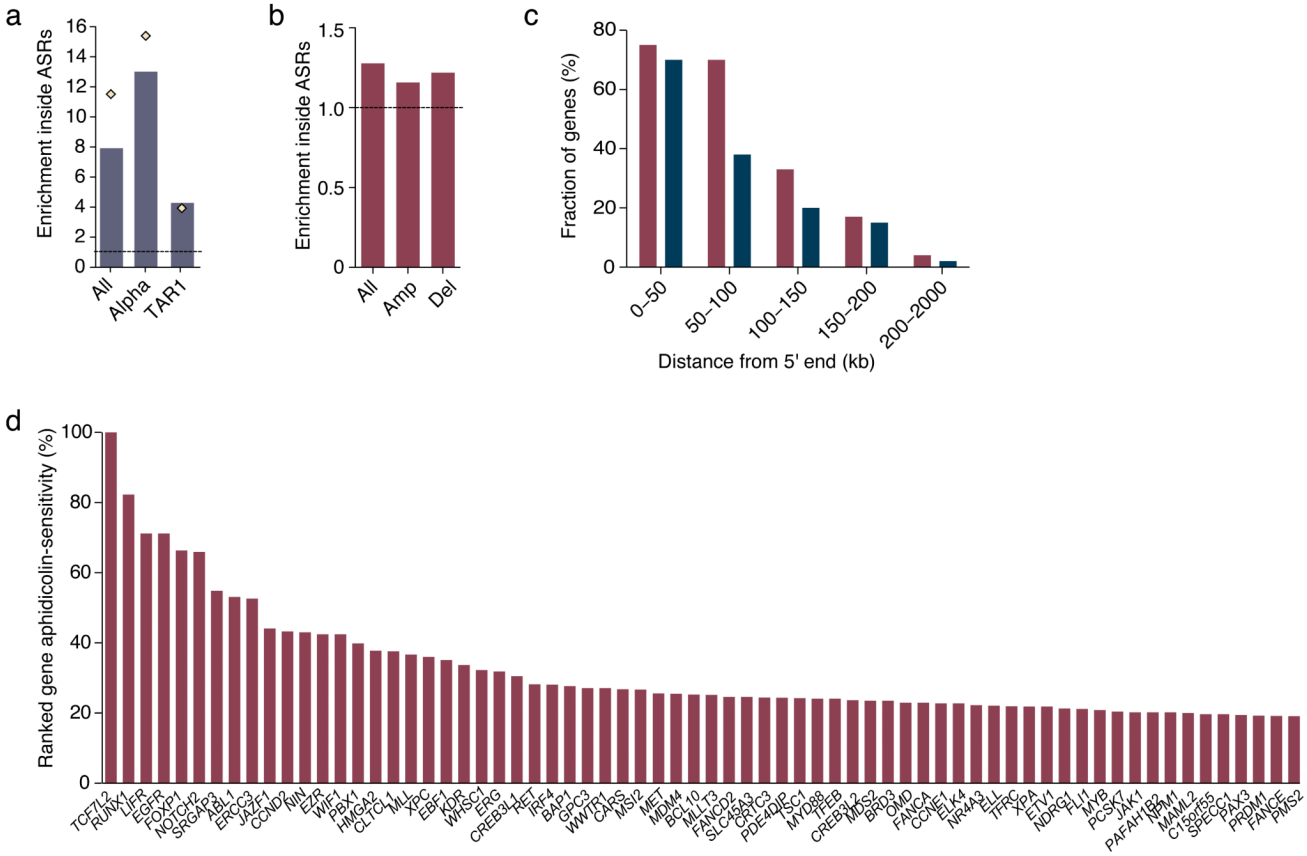


Figure 4.

Biological characterization of ASRs. (**a**) Satellite repeats significantly enriched within 48 mappable kb ASRs in comparison to the rest of the genome. Repeat names follow the nomenclature in RepeatMasker [25]. Bars: enrichments calculated based on (A1 + A2 + A3 + A4) vs. (C1 + C2 + C3 + C4) pooled samples. Diamonds: enrichments calculated based on (A1 + A2) vs. (C1 + C2) pooled samples. Dashed lines represent average genome-wide enrichment. (**b**) Significant enrichment of cancer-associated somatic copy number alterations. All: all alterations. Amp: amplifications. Del: deletions. Dashed lines represent average genome-wide enrichment. (**c**) Percentage of cancer (red) and non-cancer (blue) genes containing the center of a 48 mappable kb ASR within 2 Mb downstream from the 5′ end. (**d**) Ranking of aphidicolin-sensitive cancer-associated genes by decreasing sensitivity, expressed as percentage of the most sensitive gene on the left.